

Rich Coresets For Constrained Linear Regression

Christos Boutsidis

Mathematical Sciences Department
IBM T.J. Watson Research Center
cboutsi@us.ibm.com

Petros Drineas

Computer Science Department
Rensselaer Polytechnic Institute
drinep@cs.rpi.edu

Malik Magdon Ismail

Computer Science Department
Rensselaer Polytechnic Institute
magdon@cs.rpi.edu

February 17, 2012

Abstract

A rich coreset is a subset of the data which contains nearly all the essential information. We give deterministic, low order polynomial-time algorithms to construct rich coresets for simple and multiple response linear regression, together with lower bounds indicating that there is not much room for improvement upon our results.

1 Introduction

Linear regression is an important technique in data analysis (Seber and Lee, 1977). Research in the area ranges from numerical techniques (A. Björck, 1996) to robustness of the prediction error to noise (e.g. using feature selection (Guyon and Elisseeff, 2003)).

Is it possible to efficiently identify a *small* subset of the data that contains all the essential information of a learning problem? Such a subset is called a “rich” coreset. We show that the answer is yes, for linear regression. Such a rich coreset is analogous to the support vectors in support vector machines (Cristianini and Shawe-Taylor, 2000). Such rich coresets contain the meaningful or important points in the data and can be used to find good approximate solutions to the full problem by solving a (much) smaller problem. When the constraints are complex (e.g. non-convex constraints), solving a much smaller regression problem could be a significant saving (Gao, 2007).

We present coreset constructions for constrained regression (both simple and multiple response), as well as lower bounds for the size of “rich” coresets. In addition to potential computational savings, a rich coreset identifies the important core of a machine learning problem and is of considerable interest in applications with huge data where incremental approaches are necessary (eg. chunking) and applications where data is distributed and bandwidth is costly (hence communicating only the essential data is imperative).

Our first contribution is a deterministic, polynomial-time algorithm for constructing a rich coreset for arbitrarily constrained linear regression. Let k be the “effective dimension” of the data and let $\epsilon > 0$ be the desired accuracy parameter. Our algorithm constructs a rich coreset of size $O(k/\epsilon^2)$, which achieves a $(1 + \epsilon)$ -relative error performance guarantee. In other words, solving the regression problem on the coreset results in a solution which fits *all* the data with an error which is at most $(1 + \epsilon)$ worse than the best possible fit to all the data. We extend our results to the setting of multiple response regression using more

sophisticated techniques. Underlying our proofs are two sparsification tools from linear algebra (Batson et al., 2009; Boutsidis et al., 2011), which may be of general interest to the machine learning community.

1.1 Problem Setup

Assume the usual setting with n data points $(\mathbf{z}_1, y_1), \dots, (\mathbf{z}_n, y_n)$; $\mathbf{z}_i \in \mathbb{R}^d$ are feature vectors (which could have been obtained by applying a non-linear feature transform to raw data) and $y_i \in \mathbb{R}$ are targets (responses). The linear regression problem asks to determine a vector $\mathbf{x}_{opt} \in \mathcal{D} \subseteq \mathbb{R}^d$ that minimizes

$$\mathcal{E}(\mathbf{x}) = \sum_{i=1}^n w_i (\mathbf{z}_i^T \cdot \mathbf{x} - y_i)^2$$

over $\mathbf{x} \in \mathcal{D}$, where w_i are positive weights. So, $\mathcal{E}(\mathbf{x}_{opt}) \leq \mathcal{E}(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{D}$. The domain \mathcal{D} represents the constraints on the solution, e.g., in non-negative least squares (NNLS) (Lawson and Hanson, 1974; Bellavia et al., 2006), $\mathcal{D} = \mathbb{R}_+^d$, the nonnegative orthant. Our results hold for arbitrary \mathcal{D} .

A *coreset* of size $r < n$ is a subset of the data, $(\mathbf{z}_{i_1}, y_{i_1}), \dots, (\mathbf{z}_{i_r}, y_{i_r})$. The coreset regression problem considers the squared error on the coreset with a, possibly different, set of weights $s_j > 0$,

$$\tilde{\mathcal{E}}(\mathbf{x}) = \sum_{j=1}^r s_j (\mathbf{z}_{i_j}^T \cdot \mathbf{x} - y_{i_j})^2.$$

Suppose that $\tilde{\mathcal{E}}$ is minimized at $\tilde{\mathbf{x}}_{opt}$, so $\tilde{\mathcal{E}}(\tilde{\mathbf{x}}_{opt}) \leq \tilde{\mathcal{E}}(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{D}$. The coreset is *rich* if, for some set of weights s_j , $\tilde{\mathbf{x}}_{opt}$ is nearly as good as \mathbf{x}_{opt} for the *original regression problem* on all the data, i.e., for some small $\epsilon > 0$,

$$\mathcal{E}(\mathbf{x}_{opt}) \leq \mathcal{E}(\tilde{\mathbf{x}}_{opt}) \leq (1 + \epsilon) \mathcal{E}(\mathbf{x}_{opt}).$$

The algorithm which constructs the coreset should also provide the weights s_j . For the remainder of the paper, we switch to an equivalent matrix formulation of the problem. We give some linear algebra background in the Appendix.

Matrix Formulation. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be the *data matrix* whose rows are the weighted data points $\sqrt{w_i} \mathbf{z}_i^T$ and $\mathbf{b} \in \mathbb{R}^n$ is the similarly weighted target vector, $b_i = \sqrt{w_i} y_i$. The effective dimension of the data can be measured by the rank of \mathbf{A} ; let $k = \text{rank}(\mathbf{A})$. Our results hold for arbitrary d , however, in most applications, $n \gg d$ and $\text{rank}(\mathbf{A}) \approx d$. We can rewrite the squared error as $\mathcal{E}(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, so,

$$\mathbf{x}_{opt} = \underset{\mathbf{x} \in \mathcal{D}}{\text{argmin}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2.$$

A *coreset* of size $r < n$ is a subset $\mathbf{C} \in \mathbb{R}^{r \times d}$ of the rows of \mathbf{A} and the corresponding elements $\mathbf{b}_c \in \mathbb{R}^r$ of \mathbf{b} . Let $\mathbf{D} \in \mathbb{R}^{r \times r}$ be a positive diagonal matrix for the coreset regression (the weights s_j of the coreset regression will depend on \mathbf{D}). The weighted squared error on the coreset is given by $\tilde{\mathcal{E}}(\mathbf{x}) = \|\mathbf{D}(\mathbf{C}\mathbf{x} - \mathbf{b}_c)\|_2^2$, so the coreset regression seeks $\tilde{\mathbf{x}}_{opt}$ defined by

$$\tilde{\mathbf{x}}_{opt} = \underset{\mathbf{x} \in \mathcal{D}}{\text{argmin}} \|\mathbf{D}(\mathbf{C}\mathbf{x} - \mathbf{b}_c)\|_2^2.$$

We say that the coreset is $(1 + \epsilon)$ -rich if the solution obtained by fitting the coreset data can fit all the data almost optimally. Formally,

$$\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 \leq \|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \leq (1 + \epsilon) \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

1.2 Our contributions

Constrained Linear Regression (Section 2). Our main result for constrained simple regression is Theorem 1, which describes a deterministic polynomial time algorithm that constructs a $(1+\epsilon)$ -rich coreset of size $O(k/\epsilon^2)$. Prior to our work, the best result achieving comparable relative error performance guarantees is Theorem 1 of (Boutsidis and Drineas, 2009) for constrained regression, and the work of (Drineas et al., 2006) for unconstrained regression. Both of these prior results construct coresets of size $O(k \log k/\epsilon^2)$ and they are randomized, so, with some probability, the fit on all the data can be arbitrarily bad (despite the coreset being a logarithmic factor larger). Our methods have comparable, low order polynomial running times and provide *deterministic* guarantees. The results in (Drineas et al., 2006) and (Boutsidis and Drineas, 2009) are achieved using the matrix concentration results in (Rudelson and Vershynin, 2007). However, these concentration bounds break unless the coreset size is at least $\Omega(k \log(k)/\epsilon^2)$. We give the first algorithms that break the $k \log k$ barrier.

We extend our results to multiple response regression, where the target is a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ with $\omega \geq 1$. Each column of \mathbf{B} is a separate target (or response) that we wish to predict. We seek to minimize $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|$ over all $\mathbf{X} \in \mathcal{D} \subseteq \mathbb{R}^{d \times \omega}$, and some matrix norm $\|\cdot\|$. Multiple response regression has numerous applications, but is perhaps most common in multivariate time series analysis; see for example (Hamilton, 1994; Breiman and Friedman, 1997). To illustrate, consider prediction of time series data: let $\mathbf{Z} \in \mathbb{R}^{(n+1) \times d}$ be a set of d time series, where each column is a time series with $n+1$ time steps; we wish to predict time step $t+1$ from time step t . Let \mathbf{A} contain the first n rows of \mathbf{Z} and let \mathbf{B} contain the last n rows. Then, we seek \mathbf{X} that minimizes $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|$, which is exactly the multiple response regression problem. In our work, we focus on the spectral norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_F$, the two most common norms in matrix analysis.

Multi-Objective Regression (Section 3.1). An important variant of multiple regression is the so-called multi-objective regression. Let $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_\omega] \in \mathbb{R}^{n \times \omega}$, where we explicitly identify each column in \mathbf{B} as a target response \mathbf{b}_j . We seek to *simultaneously* fit multiple target vectors with the *same* \mathbf{x} , i.e. to simultaneously minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}_j\|_2^2$ where $j \in \{1, 2, \dots, \omega\}$. This is common when the goal is to trade off different quality criteria simultaneously. Writing $\mathbf{X} = [\mathbf{x}, \mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$ (ω copies of \mathbf{x}), we consider minimizing $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F$, which is equivalent to multiple regression with a strong constraint on \mathbf{X} . We present results for coreset constructions for the Frobenius-norm multi-objective regression problem in Theorem 4, which describes a deterministic algorithm to construct $(1+\epsilon)$ -rich coresets of size $O(k/\epsilon^2)$. Theorem 4 emerges by applying Theorem 1 after converting the Frobenius-norm multi-objective regression problem to a simple response regression problem.

Arbitrarily-Constrained Multiple-Response Regression (Section 3.2). Using the same approach, converting the problem to a single response regression, we construct a $(1+\epsilon)$ -rich coreset for Frobenius-norm arbitrarily-constrained regression in Section 3.2. The coreset size here is $O(k\omega/\epsilon^2)$.

Unconstrained Multiple-Response Regression (Section 4). In Section 4, we consider rich coresets for unconstrained multiple regression for both the spectral and Frobenius norms. The sizes of the coresets are smaller than the constrained case, and our main results are presented in Theorems 6 and 7. Theorem 6 presents a $(2+\epsilon)$ -rich coreset of size $O((k+\omega)/\epsilon^2)$ for spectral norm regression, while Theorem 7 presents a $(2+\epsilon)$ -rich coreset of size $O(k/\epsilon^2)$ for Frobenius norm regression.

Lower Bounds (Section 5). Finally, in Section 5, we present lower bounds on coreset sizes. In the single response regression setting, we note that our algorithms need to look at the target vector \mathbf{b} . We show that this is unavoidable, by arguing that no \mathbf{b} -agnostic deterministic coreset construction algorithm

can construct rich coresets which are small (Theorem 11). We also present similar results for \mathbf{b} -agnostic randomized coreset constructions (Theorem 12). Having shown that we cannot (in general) be \mathbf{b} -agnostic, we present lower bounds on the size of rich coresets for spectral and Frobenius norm multiple response regression that apply in the non \mathbf{b} -agnostic setting (Theorems 13 and 14).

2 Constrained Linear Regression

We define constrained linear regression as follows: given $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{b} \in \mathbb{R}^n$, and $\mathcal{D} \subseteq \mathbb{R}^d$, we seek $\mathbf{x}_{opt} \in \mathcal{D}$ for which $\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 \leq \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, for all $\mathbf{x} \in \mathcal{D}$ (the domain \mathcal{D} represents the constraints on \mathbf{x} and can be arbitrary). To construct a coreset $\mathbf{C} \in \mathbb{R}^{r \times d}$ (i.e., \mathbf{C} consists of a few rows of \mathbf{A}) and $\mathbf{b}_c \in \mathbb{R}^r$ (i.e., \mathbf{b}_c consists of a few elements of \mathbf{b}), we introduce *sampling* and *rescaling* matrices \mathbf{S} and \mathbf{D} respectively. More specifically, we define the *row-sampling matrix* $\mathbf{S} \in \mathbb{R}^{r \times n}$ whose rows are basis vectors $\mathbf{e}_{i_1}^T, \dots, \mathbf{e}_{i_r}^T$. Our coreset \mathbf{C} is now equal to $\mathbf{S}\mathbf{A}$; clearly, \mathbf{C} is a matrix whose rows are the rows of \mathbf{A} corresponding to indices i_1, \dots, i_r . Similarly, $\mathbf{b}_c = \mathbf{S}\mathbf{b}$ contains the corresponding elements of the target vector. Next, let $\mathbf{D} \in \mathbb{R}^{r \times r}$ be a positive diagonal rescaling matrix and define the \mathbf{D} -weighted regression problem on the coreset as follows:

$$\tilde{\mathbf{x}}_{opt} = \underset{\mathbf{x} \in \mathcal{D}}{\operatorname{argmin}} \|\mathbf{D}(\mathbf{C}\mathbf{x} - \mathbf{b}_c)\|_2^2 = \underset{\mathbf{x} \in \mathcal{D}}{\operatorname{argmin}} \|\mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{x} - \mathbf{b})\|_2^2. \quad (1)$$

In the above, the operator $\mathbf{D}\mathbf{S}$ first samples and then rescales rows of \mathbf{A} and \mathbf{b} . Theorem 1 is the main result in this section and presents a deterministic algorithm to select a rich coreset by constructing the matrices \mathbf{D} and \mathbf{S} . (All algorithms are given in the Appendix.)

Theorem 1. *Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{b} \in \mathbb{R}^n$, and $\mathcal{D} \subseteq \mathbb{R}^d$, Algorithm 1 constructs matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ (for any $r > k + 1$) such that $\tilde{\mathbf{x}}_{opt}$ of eqn. (1) satisfies*

$$\frac{\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2}{\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2} \leq \frac{r + k + 1 + 2\sqrt{r(k+1)}}{r + k + 1 - 2\sqrt{r(k+1)}} = 1 + 4\sqrt{\frac{k}{r}} + o\left(\sqrt{k/r}\right).$$

The running time of the proposed algorithm is $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}]}) + O(rnk^2)$, where $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}]})$ is the time needed to compute the left singular vectors of the matrix $[\mathbf{A}, \mathbf{b}] \in \mathbb{R}^{n \times (d+1)}$.

For any $0 < \epsilon < 1$, we can set $r = k/\epsilon^2$ to get an approximation ratio roughly equal to $1 + 4\epsilon$. This result considerably improves the result in (Boutsidis and Drineas, 2009), which needs $r = O(k \log(k)/\epsilon^2)$ to achieve the same approximation. Additionally, our bound is deterministic, whereas the bound in (Boutsidis and Drineas, 2009) fails with constant probability. (Boutsidis and Drineas, 2009) requires an SVD computation in the first step, so its running time is comparable to ours.

In order to prove the above theorem, we need a linear algebraic sparsification result from (Batson et al., 2009), which we restate using our notation.

Lemma 2 (Single-set Spectral Sparsification (Batson et al., 2009)). *Given $\mathbf{U} \in \mathbb{R}^{n \times \ell}$ satisfying $\mathbf{U}^T \mathbf{U} = \mathbf{I}_\ell$ and $r > \ell$, we can deterministically construct sampling and rescaling matrices \mathbf{S} and \mathbf{D} such that, for all $\mathbf{y} \in \mathbb{R}^\ell$:*

$$\left(1 - \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}\|_2^2 \leq \|\mathbf{D}\mathbf{S}\mathbf{U}\mathbf{y}\|_2^2 \leq \left(1 + \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}\|_2^2.$$

The algorithm runs in $O(rn\ell^2)$ time and we denote it as $[\mathbf{D}, \mathbf{S}] = \text{SimpleSampling}(\mathbf{U}, r)$.

Proof. (of Theorem 1) Let $\mathbf{Y} = [\mathbf{A}, \mathbf{b}] \in \mathbb{R}^{n \times (d+1)}$ and compute its SVD: $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T$. Let ℓ be the rank of \mathbf{Y} ($\ell \leq k+1$, since $\text{rank}(\mathbf{A}) = k$) and note that $\mathbf{U} \in \mathbb{R}^{n \times \ell}$, $\Sigma \in \mathbb{R}^{\ell \times \ell}$, and $\mathbf{V} \in \mathbb{R}^{(d+1) \times \ell}$. Let $[\mathbf{D}, \mathbf{S}] = \text{SimpleSampling}(\mathbf{U}, r)$ and define $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^\ell$ as follows:

$$\mathbf{y}_1 = \Sigma\mathbf{V}^T \begin{bmatrix} \mathbf{x}_{opt} \\ -1 \end{bmatrix}, \quad \text{and} \quad \mathbf{y}_2 = \begin{bmatrix} \tilde{\mathbf{x}}_{opt} \\ -1 \end{bmatrix}.$$

Note that $\mathbf{U}\mathbf{y}_1 = \mathbf{A}\mathbf{x}_{opt} - \mathbf{b}$, $\mathbf{U}\mathbf{y}_2 = \mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}$, $\mathbf{DSU}\mathbf{y}_1 = \mathbf{DS}(\mathbf{A}\mathbf{x}_{opt} - \mathbf{b})$, and $\mathbf{DSU}\mathbf{y}_2 = \mathbf{DS}(\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b})$. We will bound $\|\mathbf{U}\mathbf{y}_2\|$ in terms of $\|\mathbf{U}\mathbf{y}_1\|$:

$$\left(1 - \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}_2\|_2^2 \stackrel{(a)}{\leq} \|\mathbf{DSU}\mathbf{y}_2\|_2^2 \stackrel{(b)}{\leq} \|\mathbf{DSU}\mathbf{y}_1\|_2^2 \stackrel{(c)}{\leq} \left(1 + \sqrt{\ell/r}\right)^2 \|\mathbf{U}\mathbf{y}_1\|_2^2.$$

(a) and (c) follow from Lemma 2; (b) follows from the optimality of $\tilde{\mathbf{x}}_{opt}$ for the coreset regression in eqn. (1). Using $\ell \leq k+1$ and manipulating the above expression concludes the proof of the theorem. The running time of the algorithm is equal to the time needed to compute \mathbf{U} and the time needed to run the algorithm of Lemma 2 with $\ell \leq k+1$. \blacksquare

3 Constrained Multiple-Response Regression

Constrained multiple-response regression in the Frobenius norm can be reduced to simple regression. So, we can apply the results of the previous section to this setting.

3.1 Multi-Objective Regression

The task is to minimize, over all $\mathbf{x} \in \mathcal{D}$, the Frobenius-norm error $\|\mathbf{A}[\mathbf{x}, \dots, \mathbf{x}] - \mathbf{B}\|_F^2$. Let $\mathbf{b}_{avg} = \frac{1}{\omega}\mathbf{B}\mathbf{1}_\omega$ (here $\mathbf{1}_\omega$ is a vector of all ones and thus \mathbf{b}_{avg} is the average of the columns in \mathbf{B}). Recall that $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and let $\mathbf{X} = [\mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$.

Lemma 3. For $\mathbf{X} = [\mathbf{x}, \dots, \mathbf{x}] \in \mathbb{R}^{d \times \omega}$, $\|\mathbf{AX} - \mathbf{B}\|_F^2 = \omega\|\mathbf{Ax} - \mathbf{b}_{avg}\|_2^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|_2^2$.

In the above $\mathbf{B}^{(i)}$ denotes the i -th column of \mathbf{B} as a column vector. Note that the second term in Lemma 3 does not depend on \mathbf{x} and thus the generalized multi-objective regression can be reduced to simple regression on \mathbf{A} and \mathbf{b}_{avg} . Using Theorem 1, we can get a coreset: let $\tilde{\mathbf{x}}_{opt}$ minimize $\|\mathbf{DS}(\mathbf{Ax} - \mathbf{b}_{avg})\|_2$, where \mathbf{S} and \mathbf{D} are obtained via Theorem 1 applied to \mathbf{A} and \mathbf{b}_{avg} . If $\tilde{\mathbf{X}}_{opt} = [\tilde{\mathbf{x}}_{opt}, \dots, \tilde{\mathbf{x}}_{opt}]$, then, by Lemma 3, $\tilde{\mathbf{X}}_{opt}$ minimizes $\|\mathbf{DS}(\mathbf{AX} - \mathbf{B})\|_F$. Similarly, if \mathbf{x}_{opt} minimizes $\|\mathbf{Ax} - \mathbf{b}_{avg}\|_2$ and $\mathbf{X}_{opt} = [\mathbf{x}_{opt}, \dots, \mathbf{x}_{opt}]$, then \mathbf{X}_{opt} minimizes $\|\mathbf{AX} - \mathbf{B}\|_F$. Theorem 4 says that $\tilde{\mathbf{X}}_{opt}$ approximates \mathbf{X}_{opt} (the proof is in the appendix).

Theorem 4. Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, we can construct matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$ (for any $r > k+1$) such that the matrix $\tilde{\mathbf{X}}_{opt} = [\tilde{\mathbf{x}}_{opt}, \dots, \tilde{\mathbf{x}}_{opt}]$ that minimizes $\|\mathbf{DS}(\mathbf{AX} - \mathbf{B})\|_F$ over all matrices $\mathbf{X} = [\mathbf{x}, \mathbf{x}, \dots, \mathbf{x}]$ satisfies:

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_F^2 \leq \left(1 + O\left(\sqrt{k/r}\right)\right) \|\mathbf{AX}_{opt} - \mathbf{B}\|_F^2.$$

The run time of the proposed algorithm is $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}_{avg}]}) + O(n\omega + rnk^2)$, where $T(\mathbf{U}_{[\mathbf{A}, \mathbf{b}_{avg}]})$ is the time needed to compute the left singular vectors of the matrix $[\mathbf{A}, \mathbf{b}_{avg}] \in \mathbb{R}^{n \times (d+1)}$.

We note that the coreset size depends only on the rank of \mathbf{A} and not on the size of \mathbf{B} .

3.2 Arbitrarily-Constrained Multiple-Response Regression

Multi-objective regression is a special case of constrained multiple-response regression for which we can efficiently obtain the coresets. In the general case, the problem still reduces to simple regression, but the coresets are now larger. We wish to minimize $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F$ over $\mathbf{X} \in \mathcal{D} \subseteq \mathbb{R}^{d \times \omega}$. Since $\mathbb{R}^{d \times \omega}$ is isomorphic to $\mathbb{R}^{d\omega}$, we can view $\mathbf{X} \in \mathbb{R}^{d \times \omega}$ as a “stretched out” vector $\hat{\mathbf{X}} \in \mathbb{R}^{d\omega}$; corresponding to the domain \mathcal{D} is the domain $\hat{\mathcal{D}} \subseteq \mathbb{R}^{d\omega}$. Similarly, we can stretch out $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ to $\hat{\mathbf{B}} \in \mathbb{R}^{n\omega}$. To complete the transformation to simple linear regression, we build a transformed block-diagonal data matrix $\hat{\mathbf{A}}$ from \mathbf{A} , by repeating ω copies of \mathbf{A} along the diagonal:

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & & & \\ & \mathbf{A} & & \\ & & \ddots & \\ & & & \mathbf{A} \end{bmatrix} \in \mathbb{R}^{n\omega \times d\omega}, \quad \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \\ \vdots \\ \mathbf{X}^{(\omega)} \end{bmatrix} \in \mathbb{R}^{d\omega}, \quad \hat{\mathbf{B}} = \begin{bmatrix} \mathbf{B}^{(1)} \\ \mathbf{B}^{(2)} \\ \vdots \\ \mathbf{B}^{(\omega)} \end{bmatrix} \in \mathbb{R}^{n\omega}$$

Lemma 5. For all \mathbf{A} , \mathbf{X} and \mathbf{B} of appropriate dimensions, $\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_F^2 = \|\hat{\mathbf{A}}\hat{\mathbf{X}} - \hat{\mathbf{B}}\|_2^2$.

Theorem 1 gives us coresets for this equivalent regression. Note that $\text{rank}(\hat{\mathbf{A}}) \leq \omega \cdot \text{rank}(\mathbf{A})$. The coreset will identify the important rows of \mathbf{A} (the same row may get identified multiple times as different rows of $\hat{\mathbf{A}}$), and the important *elements* of \mathbf{B} , because the entries in $\hat{\mathbf{B}}$ are elements of \mathbf{B} , not rows of \mathbf{B} . Let $\hat{\mathbf{X}}_{opt}$ be the solution constructed from the coreset, which minimizes $\|\hat{\mathbf{A}}\hat{\mathbf{X}} - \hat{\mathbf{B}}\|$ over $\hat{\mathbf{X}} \in \hat{\mathcal{D}}$, and let $\tilde{\mathbf{X}}_{opt} \in \mathcal{D}$ be the corresponding solution in the original domain \mathcal{D} . If r is the size of the coreset and $\text{rank}(\mathbf{A}) = k$, then, by Theorem 1,

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_F^2 \leq \left(1 + O\left(\sqrt{k\omega/r}\right)\right) \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_F^2.$$

So, for the approximation ratio to be $1 + O(\epsilon)$, we set $r = O(k\omega/\epsilon^2)$. The running time would involve the time needed to compute the SVD of $[\hat{\mathbf{A}}, \hat{\mathbf{B}}]$.

Notice that the coresets are large and somewhat costly to compute and they only work for the Frobenius norm. In the next section, using more sophisticated techniques, we will get smaller coresets for unconstrained regression in both the Frobenius and spectral norms.

4 Unconstrained Multiple-Response Regression

Consider the following problem: given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank exactly k and a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, we seek to identify the matrix $\mathbf{X}_{opt} \in \mathbb{R}^{d \times \omega}$ that minimizes ($\xi = 2$ and $\xi = F$)

$$\mathbf{X}_{opt} = \arg \min_{\mathbf{X} \in \mathbb{R}^{d \times \omega}} \|\mathbf{A}\mathbf{X} - \mathbf{B}\|_\xi^2.$$

We can compute \mathbf{X}_{opt} via the pseudoinverse of \mathbf{A} , namely $\mathbf{X}_{opt} = \mathbf{A}^+ \mathbf{B}$. If \mathbf{S} and \mathbf{D} are sampling and rescaling matrices respectively, then the *coreset* regression problem is:

$$\tilde{\mathbf{X}}_{opt} = \arg \min_{\mathbf{X} \in \mathbb{R}^{d \times \omega}} \|\mathbf{D}\mathbf{S}\mathbf{A}\mathbf{X} - \mathbf{D}\mathbf{S}\mathbf{B}\|_\xi^2. \quad (2)$$

The solution of the *coreset* regression problem is $\tilde{\mathbf{X}}_{opt} = (\mathbf{D}\mathbf{S}\mathbf{A})^+ \mathbf{D}\mathbf{S}\mathbf{B}$. The main results in this section are presented in Theorems 6 and 7.

Theorem 6 (Spectral norm). *Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank exactly k and a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, Algorithm 2 deterministically constructs matrices \mathbf{S} and \mathbf{D} such that solving the problem of eqn. (2) satisfies (for any r such that $k + 1 < r \leq n$):*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_2^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2 + \left(\frac{1 + \sqrt{\omega/r}}{1 - \sqrt{k/r}} \right)^2 \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2.$$

The running time of the proposed algorithm is $T(\mathbf{U}_{\mathbf{A}}) + O(rn(k^2 + \omega^2))$, where $T(\mathbf{U}_{\mathbf{A}})$ is the time needed to compute the left singular vectors of \mathbf{A} .

Asymptotically, for large ω , the approximation ratio of the above theorem is $O(\omega/r)$. We will argue that this is nearly optimal by providing a matching lower bound in Theorem 13.

Theorem 7 (Frobenius norm). *Given matrices $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, Algorithm 3 deterministically constructs a sampling matrix \mathbf{S} and a rescaling matrix \mathbf{D} such that solving the problem of eqn. (2) satisfies (for any r such that $k + 1 < r \leq n$):*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_F^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_F^2 + \frac{1}{(1 - \sqrt{k/r})^2} \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_F^2.$$

The running time of the proposed algorithm is $T(\mathbf{U}_{\mathbf{A}}) + O(rnk^2)$, where $T(\mathbf{U}_{\mathbf{A}})$ is the time needed to compute the left singular vectors of \mathbf{A} .

The approximation ratio in the above theorem is $2 + O(\sqrt{k/r})$. In Theorem 14, we will give a lower bound for the approximation ratio which is $1 + \Omega(k/r)$. We conjecture that our lower bound can be achieved, perhaps by a more sophisticated algorithm.

Finally, we note that the **B-agnostic** randomized construction of Drineas et al. (2008) achieves a $(1 + \epsilon)$ approximation ratio using a significantly larger coreset, $r = O(k \log(k)/\epsilon^2)$. Importantly, they *do not need any access to \mathbf{B}* in order to construct the coreset, whereas our approach constructs coresets by carefully choosing important data points with respect to the particular target response matrix \mathbf{B} . We will also discuss **B-agnostic** algorithms in Section 4.2 (Theorem 10) and we will present matching lower bounds in Section 5.

4.1 Proofs of Theorems 6 and 7

We will make heavy use of facts from Section A. We start with a few simple lemmas.

Lemma 8. *Let $\mathbf{E} = \mathbf{A}\mathbf{X}_{opt} - \mathbf{B}$ be the regression residual. Then, $\text{rank}(\mathbf{E}) \leq \min\{\omega, n - k\}$.*

Proof. Using our notation, $\mathbf{A}\mathbf{X}_{opt} - \mathbf{B} = (\mathbf{I}_n - \mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^T)\mathbf{B} = \mathbf{U}_{\mathbf{A}}^\perp (\mathbf{U}_{\mathbf{A}}^\perp)^T \mathbf{B}$. To conclude notice that $\text{rank}(\mathbf{X}\mathbf{Y}) \leq \min\{\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y})\}$. \blacksquare

We now give our main tool for obtaining approximation guarantees for coreset regression. The proof is deferred to the appendix.

Lemma 9. *Assume that the rank of the matrix $\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{A}} \in \mathbb{R}^{r \times k}$ is equal to k (i.e., the matrix has full rank). Then, for $\xi = 2, F$,*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_\xi^2 \leq \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_\xi^2 + \|(\mathbf{D}\mathbf{S}\mathbf{U}_{\mathbf{A}})^+ \mathbf{D}\mathbf{S}(\mathbf{A}\mathbf{X}_{opt} - \mathbf{B})\|_\xi^2.$$

This lemma provides a framework for coresets construction: all we need are sampling and rescaling matrices \mathbf{S} and \mathbf{D} , such that $\text{rank}(\mathbf{DSU}_\mathbf{A}) = k$ and $\|(\mathbf{DSU}_\mathbf{A})^+ \mathbf{DS}(\mathbf{AX}_{opt} - \mathbf{B})\|_\xi^2$ is small. The final ingredients for the proofs of Theorems 6 and 7 are two matrix sparsification results, Lemmas 16 and 17 in the Appendix.

Proof. (of Theorem 6) Theorem 6 follows from Lemmas 9 and 16. First, compute the SVD of \mathbf{A} to obtain $\mathbf{U}_\mathbf{A}$, and let $\mathbf{E} = \mathbf{AX}_{opt} - \mathbf{B} = \mathbf{U}_\mathbf{A} \mathbf{U}_\mathbf{A}^T \mathbf{B} - \mathbf{B}$. Next, run the algorithm of Lemma 16 to obtain $[\Omega, \mathbf{S}] = \text{MultipleSpectralSampling}(\mathbf{U}_\mathbf{A}, \mathbf{E}, r)$. This algorithm will run in time $T_{SVD}(\mathbf{E}) + O(rn(k^2 + \rho_\mathbf{E}^2))$, where k is the rank of $\mathbf{U}_\mathbf{A}$ and \mathbf{A} . The total running time of the algorithm is $T(\mathbf{U}_\mathbf{A}) + T_{SVD}(\mathbf{E}) + O(rn(k^2 + \rho_\mathbf{E}^2)) = T(\mathbf{U}_\mathbf{A}) + O(rn(k^2 + \omega^2))$.

Lemma 16 guarantees that \mathbf{D} and \mathbf{S} satisfy the rank assumption of Lemma 9. To conclude the proof, we bound the second term of Lemma 9, using the bounds of Lemma 16 and $\rho_\mathbf{E} \leq \min\{\omega, n - k\} \leq \omega$:

$$\begin{aligned} \|(\mathbf{DSU}_\mathbf{A})^+ \mathbf{DS}(\mathbf{AX}_{opt} - \mathbf{B})\|_2^2 &\leq \|(\mathbf{DSU}_\mathbf{A})^+\|_2^2 \|\mathbf{DS}(\mathbf{AX}_{opt} - \mathbf{B})\|_2^2 \\ &\leq \left(1 + \sqrt{\omega/r}\right)^2 \left(1 - \sqrt{k/r}\right)^{-2} \|\mathbf{AX}_{opt} - \mathbf{B}\|_2^2. \end{aligned}$$

■

Proof. (of Theorem 7) Similar to Theorem 6, using Lemma 17 instead of Lemma 16. ■

4.2 B-Agnostic Coresets Construction

All the coresets construction algorithms that we presented so far carefully construct the coreset using knowledge of the response vector. If the algorithm does not need knowledge of \mathbf{B} to construct the coreset, and yet can provide an approximation guarantee for every \mathbf{B} , then the algorithm is \mathbf{B} -agnostic. A \mathbf{B} -agnostic coreset construction algorithm is appealing because the coreset, as specified by the sampling and rescaling matrices \mathbf{S} and \mathbf{D} , can be computed off-line and applied to *any* \mathbf{B} . We briefly digress to show how our methods can be extended to develop \mathbf{B} -agnostic coreset constructions.

Theorem 10 (B-Agnostic Coresets). *Given a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank exactly k and a matrix $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, there exists an algorithm to deterministically construct a sampling matrix \mathbf{S} and a rescaling matrix \mathbf{D} such that for any $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, the matrix $\tilde{\mathbf{X}}_{opt}$ that solves the problem of eqn. (2) satisfies (for any r such that $k < r \leq n$):*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_\xi^2 \leq \|\mathbf{AX}_{opt} - \mathbf{B}\|_\xi^2 + \left(\frac{1 + \sqrt{n/r}}{1 - \sqrt{k/r}}\right)^2 \|\mathbf{AX}_{opt} - \mathbf{B}\|_\xi^2.$$

The running time of the proposed algorithm is $T(\mathbf{U}_\mathbf{A}) + O(rnk^2)$, where $T(\mathbf{U}_\mathbf{A})$ is the time needed to compute the left singular vectors of \mathbf{A} .

Proof. The proof is similar to the proof of Theorem 6, except we now construct the sampling and rescaling matrices as $[\mathbf{D}, \mathbf{S}] = \text{MultipleSpectralSampling}(\mathbf{U}_\mathbf{A}, \mathbf{I}_n, r)$. To bound the second term in Lemma 9, we use

$$\begin{aligned} \|(\mathbf{DSU}_\mathbf{A})^+ \mathbf{DS}(\mathbf{AX}_{opt} - \mathbf{B})\|_\xi^2 &= \|(\mathbf{DSU}_\mathbf{A})^+ \mathbf{DSI}_n(\mathbf{AX}_{opt} - \mathbf{B})\|_\xi^2 \\ &\leq \|(\mathbf{DSU}_\mathbf{A})^+\|_2^2 \|\mathbf{DSI}_n\|_2^2 \|\mathbf{AX}_{opt} - \mathbf{B}\|_\xi^2, \end{aligned}$$

and the bounds of Lemma 16. ■

The above bound decreases with r and holds for any \mathbf{B} , guaranteeing a constant-factor approximation with a constant fraction of the data. The approximation ratio is $O(n/r)$, which seems quite weak. In the next section, we show that this result is tight.

5 Lower Bounds on Coreset Size

We have just seen a \mathbf{B} -agnostic coreset construction algorithm with a rather weak worst case guarantee of $O(n/r)$ approximation error. We will now show that no deterministic \mathbf{B} -agnostic coreset construction algorithm can guarantee a better error (Theorem 11).

(Drineas et al., 2008) provides another \mathbf{B} -agnostic coreset construction algorithm with $r = O(k \log(k)/\epsilon^2)$. For a fixed \mathbf{B} , the method in (Drineas et al., 2008) delivers a probabilistic bound on the approximation error. However, there are target matrices \mathbf{B} for which the bound fails by an arbitrarily large amount. The probabilistic algorithms get away with this by brushing all these (possibly large) errors into a low probability event, with respect to random choices made in the algorithm. So, in some sense, these algorithms are not \mathbf{B} -agnostic, in that they do not construct a coreset which works well *for all* \mathbf{B} with some (say) constant probability. Nevertheless, the fact that they give a constant probability of success for a fixed *but unknown* \mathbf{B} makes these algorithms interesting and useful. We will give a lower bound on the approximation ratio of such algorithms as well, for a given probability of success (Theorem 12). Finally, we will give lower bounds on the size of the coreset for the general (non-agnostic) multiple regression setting (Theorems 13 and 14).

5.1 An Impossibility Result for \mathbf{B} -Agnostic Coreset Construction

We first present the lower bound for simple regression. Recall that a coreset construction algorithm is \mathbf{b} -agnostic if it constructs a coreset without knowledge of \mathbf{b} , and then provides an approximation guarantee for *every* \mathbf{b} . We show that no coreset can work for *every* \mathbf{b} ; therefore a \mathbf{b} -agnostic coreset will be bad for some vector \mathbf{b} . In fact, there exists a matrix \mathbf{A} such that every coreset has an associated “bad” \mathbf{b} .

Theorem 11 (Deterministic \mathbf{b} -Agnostic coresets). *There exists a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ such that for every coreset $\mathbf{C} \in \mathbb{R}^{r \times d}$ of size $r \leq n$, there exists $\mathbf{b} \in \mathbb{R}^n$ (depending on \mathbf{C}) for which*

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \geq \frac{n}{r} \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

Proof. Let \mathbf{A} be any matrix with orthonormal columns whose first column is $\mathbf{1}_n/\sqrt{n}$, and consider any coreset \mathbf{C} of size r . Let $\mathbf{b} = \mathbf{1}_{\bar{\mathbf{C}}}/\sqrt{n-r}$, where $\mathbf{1}_{\bar{\mathbf{C}}}$ is the n -vector of 1’s except at the coreset locations. So for the coreset regression, $\mathbf{b}_c = \mathbf{0}$, and so $\tilde{\mathbf{x}}_{opt} = \mathbf{0}_{d \times 1}$. Therefore, $\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 = \|\mathbf{b}\|_2^2 = 1$. Let $\mathbf{P}_{\mathbf{A}}$ project onto the columns of \mathbf{A} and $\mathbf{P}_{\mathbf{A}(1)}$ project onto the first column of \mathbf{A} . The following sequence establishes the result:

$$\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\mathbf{b}\|_2^2 \leq \|(\mathbf{I} - \mathbf{P}_{\mathbf{A}(1)})\mathbf{b}\|_2^2 = \frac{r}{n}$$

■

We now consider randomized algorithms that construct a coreset without looking at \mathbf{b} (e.g. (Drineas et al., 2008)). These algorithms work for any fixed (but unknown) \mathbf{b} , and deliver a probabilistic approximation guarantee for any single fixed \mathbf{b} ; in some sense they are \mathbf{b} -agnostic. By the previous discussion, the returned coreset must fail for some \mathbf{b} , i.e., the probabilistic guarantee does not hold for all \mathbf{b} and, when it fails, it could do so with very bad error. We will now present a lower bound on the approximation accuracy of such existing randomized algorithms for coreset construction, even for a *single* \mathbf{b} .

First, we define randomized coreset construction algorithms. Let $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_{\binom{n}{r}}$ be the $\binom{n}{r}$ different coresets of size r . A randomized algorithm assigns probabilities $p_1, p_2, \dots, p_{\binom{n}{r}}$ to each coreset, and selects one according to these probabilities. The probabilities p_i may depend on \mathbf{A} . The algorithm is \mathbf{b} -agnostic if the probabilities p_i do not depend on \mathbf{b} .

Theorem 12 (Probabilistic b -Agnostic Coresets). *For any randomized \mathbf{b} -agnostic coreset construction algorithm, and any integer $0 \leq \ell \leq n - r$, there exists $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{b} \in \mathbb{R}^n$, such that, with probability at least $\binom{n-r}{\ell} / \binom{n}{\ell}$,*

$$\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \geq \frac{n}{n-\ell} \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2.$$

Proof. Let \mathbf{A} be any matrix with orthonormal columns whose first column is $\mathbf{1}_n/\sqrt{n}$, as in the proof of Theorem 11. Let \mathbf{T} be a set of size $\ell \leq n - r$. The neighborhood $N(\mathbf{T})$ is the set of coresets that have non-empty intersection with \mathbf{T} . Every coreset appears in $\binom{n}{\ell} - \binom{n-r}{\ell}$ such neighborhoods (the number of sets of size ℓ which intersect with a coreset of size r). Let $\Pr[\mathbf{T}]$ be the probability that the coreset selected by the algorithm is in $N(\mathbf{T})$; then, $\Pr[\mathbf{T}] = \sum_{\mathbf{C}_i \in N(\mathbf{T})} \Pr[\mathbf{C}_i]$. Therefore,

$$\sum_{\mathbf{T}} \Pr[\mathbf{T}] = \sum_{\mathbf{T}} \sum_{\mathbf{C}_i \in N(\mathbf{T})} \Pr[\mathbf{C}_i] = \binom{n}{\ell} - \binom{n-r}{\ell},$$

where the last equality follows because each coreset appears exactly $\binom{n}{\ell} - \binom{n-r}{\ell}$ times in the summation and $\sum_i \Pr[\mathbf{C}_i] = 1$. Thus, there is at least one set \mathbf{T}^* for which

$$\Pr[\mathbf{C} \in N(\mathbf{T}^*)] \leq \frac{\binom{n}{\ell} - \binom{n-r}{\ell}}{\binom{n}{\ell}} = 1 - \frac{\binom{n-r}{\ell}}{\binom{n}{\ell}}.$$

So with probability at least $\binom{n-r}{\ell} / \binom{n}{\ell}$, the selected coreset does not intersect with \mathbf{T}^* . Select $\mathbf{b} = \mathbf{1}_{\mathbf{T}^*}$ (the unit vector which is $1/\sqrt{\ell}$ at the indices corresponding to \mathbf{T}^*). Now, with probability at least $\binom{n-r}{\ell} / \binom{n}{\ell}$, $\tilde{\mathbf{x}}_{opt} = \mathbf{0}$, and the analysis in the proof of Theorem 11 shows that $\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}\|_2^2 \geq \frac{n}{n-\ell} \|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}\|_2^2$. ■

By Stirling's formula, after some algebra, the probability $\binom{n-r}{\ell} / \binom{n}{\ell}$ is asymptotic to $e^{-2r\ell/n}$. Setting $\ell = \Theta(n/r)$ gives a success probability that is $\Theta(1)$ (a constant), then the approximation ratio cannot be better than $1 + \Omega(1/r)$. With regard to high probability (approaching one) algorithms, consider $\ell = n \log n / 2r$ to conclude that if the success probability is at least $1 - 1/n$, the approximation ratio is no better than $1 + \log(n)/(2r - \log n)$.

5.2 Lower Bounds for Non-Agnostic Multiple Regression

For both the spectral and the Frobenius norm, we now consider non-agnostic unconstrained multiple regression, and give lower bounds for coresets of size $r > d = \text{rank}(\mathbf{A})$ (for simplicity, we set $\text{rank}(\mathbf{A}) = d$). The results are presented in Theorems 13 and 14.

Theorem 13 (Spectral Norm). *There exists $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ such that for any $r > d$ and any sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$, the solution to the coreset regression $\tilde{\mathbf{X}}_{opt} = (\mathbf{D}\mathbf{S}\mathbf{A})^+ \mathbf{D}\mathbf{S}\mathbf{B} \in \mathbb{R}^{d \times \omega}$ satisfies*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_2^2 \geq \frac{w}{r+1} \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_2^2.$$

Proof. First, we need some results from (Boutsidis et al., 2011). Boutsidis et al. (2011) exhibits a matrix $\mathbf{B} \in \mathbb{R}^{(\omega-1) \times \omega}$ such that for any sampling matrix $\mathbf{S} \in \mathbb{R}^{r \times (\omega-1)}$ and rescaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$, with $\mathbf{C} = \mathbf{D}\mathbf{S}\mathbf{B}$ (rescaled sampled coreset of \mathbf{B}),

$$\|\mathbf{B} - \Pi_{\mathbf{C},k}(\mathbf{B})\|_2^2 \geq \frac{\omega}{r+1} \|\mathbf{B} - \mathbf{B}_k\|_2^2,$$

where $\Pi_{\mathbf{C},k}(\mathbf{B})$ is the best rank- k approximation to \mathbf{B} whose rows lie in the span of all the rows in \mathbf{C} (the row-space of \mathbf{C}); and, \mathbf{B}_k is the best rank- k approximation to \mathbf{B} (which could be computed via the truncated SVD of \mathbf{B}). Actually, \mathbf{D} is irrelevant here because the row-space of \mathbf{SB} is not changed by a positive diagonal rescaling matrix \mathbf{D} .

Since $\Pi_{\mathbf{C},k}(\mathbf{B})$ is the best rank- k approximation to \mathbf{B} in the row-space of \mathbf{C} , it follows that $\|\mathbf{B} - \Pi_{\mathbf{C},k}(\mathbf{B})\|_2^2 \leq \|\mathbf{B} - \mathbf{XC}\|_2^2$ for any $\mathbf{X} \in \mathbb{R}^{(\omega-1) \times r}$ with rank at most k (because \mathbf{XC} will have rank at most k and is in the row space of \mathbf{C}). Set $\mathbf{X} = \mathbf{U}_{\mathbf{B},k}(\mathbf{DSU}_{\mathbf{B},k})^+$, where $\mathbf{U}_{\mathbf{B},k} \in \mathbb{R}^{(\omega-1) \times k}$ has k columns which are the top- k left singular vectors of \mathbf{B} . It is easy to verify that \mathbf{X} has the correct dimensions and rank at most k . Since $\mathbf{C} = \mathbf{DSB}$, we have that

$$\|\mathbf{B} - \Pi_{\mathbf{C},k}(\mathbf{B})\|_2^2 \leq \|\mathbf{B} - \mathbf{U}_{\mathbf{B},k}(\mathbf{DSU}_{\mathbf{B},k})^+ \mathbf{DSB}\|_2^2.$$

We now construct the regression problem. Let $\mathbf{A} = \mathbf{U}_{\mathbf{B},d} \in \mathbb{R}^{(\omega-1) \times d}$ (i.e., we choose $k = d$ in the above discussion and $n = \omega - 1$). Suppose a coreset construction algorithm gives sampling and rescaling matrices \mathbf{S} and \mathbf{D} , for a coreset of size r . So, the coreset regression is with $\tilde{\mathbf{A}} = \mathbf{C} = \mathbf{DSA}$ and $\tilde{\mathbf{B}} = \mathbf{DSB}$. The solution to the coreset regression is

$$\tilde{\mathbf{X}}_{opt} = \tilde{\mathbf{A}}^+ \tilde{\mathbf{B}} = \mathbf{C}^+ \mathbf{DSB} = (\mathbf{DSA})^+ \mathbf{DSB} = (\mathbf{DSU}_{\mathbf{B},d})^+ \mathbf{DSB},$$

which means that

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_2^2 = \|\mathbf{U}_{\mathbf{B},d}(\mathbf{DSU}_{\mathbf{B},d})^+ \mathbf{DSB} - \mathbf{B}\|_2^2 \geq \|\Pi_{\mathbf{C},d}(\mathbf{B}) - \mathbf{B}\|_2^2 \geq \frac{\omega}{r+1} \|\mathbf{B}_d - \mathbf{B}\|_2^2.$$

To conclude the proof, observe that $\mathbf{B}_d = \mathbf{U}_{\mathbf{B},d} \mathbf{U}_{\mathbf{B},d}^T \mathbf{B} = \mathbf{AA}^+ \mathbf{B} = \mathbf{AX}_{opt}$. ■

Theorem 14 (Frobenius Norm). *There exists $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{B} \in \mathbb{R}^{n \times \omega}$ such that for any $r > d$ and any sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{n \times r}$ and $\mathbf{D} \in \mathbb{R}^{r \times r}$, the solution to the coreset regression $\tilde{\mathbf{X}}_{opt} = (\mathbf{DSA})^+ \mathbf{DSB} \in \mathbb{R}^{d \times \omega}$ satisfies*

$$\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_F^2 \geq \left(1 + \frac{d}{r}\right) \|\mathbf{AX}_{opt} - \mathbf{B}\|_F^2.$$

Proof. The proof of this Frobenius norm lower bound follows the same argument as in the proof of Theorem 13, with $\omega/(r+1)$ replaced by $1 + d/r$, providing that there is a matrix \mathbf{B} for which $\|\mathbf{B} - \Pi_{\mathbf{C},d}(\mathbf{B})\|_F^2 \geq (1 + d/r) \|\mathbf{B} - \mathbf{B}_d\|_F^2$. Indeed, the construction of such a matrix was presented in (Boutsidis et al., 2011). ■

6 Open problems

Can one determine the minimum size of a coreset that provides a $(1 + \epsilon)$ relative-error guarantee for simple linear regression? We conjecture that $\Omega(k/\epsilon)$ is a lower bound, which will make our results almost tight. Certainly, rich coresets of size exactly k cannot be guaranteed: consider two data points $(1, 1), (-1, 1)$. The optimal regression is 0; however any coreset of size one will give non-zero regression. Is it possible to get strong guarantees on small coresets for other learning problems?

References

- A. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- J.D. Batson, D.A. Spielman, and N. Srivastava. Twice-ramanujan sparsifiers. In *Proc. 41st Annual ACM STOC*, pages 255–262, 2009.
- S. Bellavia, M. Macconi, and B. Morini. An interior point newton-like method for non-negative least squares problems with degenerate solution. *Numerical Linear Algebra with Applications*, 13:825–844, 2006.
- M.W. Berry, S.T. Dumais, and G.W. O’Brian. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.
- C. Boutsidis. *Topics in Matrix Sampling Algorithms*. PhD thesis, Rensselaer Polytechnic Institute, 2011. <http://arxiv.org/abs/1105.0709>.
- C. Boutsidis and P. Drineas. Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications*, 431(5-7):760–771, 2009.
- C. Boutsidis, P. Drineas, and M. Magdon-Ismael. Near-optimal column based matrix reconstruction. *Preprint, Available online, ArXiv*, 2011.
- L. Breiman and J. Friedman. Predicting multivariate responses in multiple linear regression. *J. Royal Stat. Soc.*, 59(1):3–54, 1997.
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.
- P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proc. SODA*, pages 1127–1136, 2006.
- P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal Matrix Analysis and Applications*, 30(2):844–881, 2008.
- D.Y. Gao. Solutions and optimality criteria to box constrained nonconvex minimization problems. *MANAGEMENT*, 3(2):293–304, 2007.
- I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- C. L. Lawson and R. J. Hanson. Solving least squares problems. *Prentice-Hall*, 1974.
- M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. of the ACM*, 54, 2007.
- G.A.F. Seber and A.J. Lee. *Linear regression analysis*. Wiley New York, 1977. ISBN 0471019674.

A Linear Algebra Background

The Singular Value Decomposition (SVD) of a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k is a decomposition $\mathbf{A} = \mathbf{U}_\mathbf{A} \Sigma_\mathbf{A} \mathbf{V}_\mathbf{A}^\top$. The singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k > 0$ are contained in the diagonal matrix $\Sigma_\mathbf{A} \in \mathbb{R}^{k \times k}$; $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$ contains the left singular vectors of \mathbf{A} ; and $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{d \times k}$ contains the right singular vectors. The SVD of \mathbf{A} can be computed in deterministic $O(nd \min\{n, d\})$ time.

The Moore-Penrose pseudo-inverse of \mathbf{A} is equal to $\mathbf{A}^+ = \mathbf{V}_\mathbf{A} \Sigma_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A}^\top$. Given an orthonormal matrix $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$, the perpendicular matrix $\mathbf{U}_\mathbf{A}^\perp \in \mathbb{R}^{n \times (n-k)}$ to $\mathbf{U}_\mathbf{A}$ satisfies: $(\mathbf{U}_\mathbf{A}^\perp)^\top \mathbf{U}_\mathbf{A}^\perp = \mathbf{I}_{n-k}$, $\mathbf{U}_\mathbf{A}^\top \mathbf{U}_\mathbf{A}^\perp = \mathbf{0}_{k \times (n-k)}$, and $\mathbf{U}_\mathbf{A} \mathbf{U}_\mathbf{A}^\top + \mathbf{U}_\mathbf{A}^\perp (\mathbf{U}_\mathbf{A}^\perp)^\top = \mathbf{I}_n$. All the singular values of both $\mathbf{U}_\mathbf{A}$ and $\mathbf{U}_\mathbf{A}^\perp$ are equal to one. Given $\mathbf{U}_\mathbf{A}$, $\mathbf{U}_\mathbf{A}^\perp$ can be computed in deterministic $O(n(n-k)^2)$ time via the QR factorization.

We remind the reader of the Frobenius and spectral matrix norms: $\|\mathbf{A}\|_\text{F}^2 = \sum_{i,j} \mathbf{A}_{ij}^2 = \sum_{i=1}^k \sigma_i^2$ and $\|\mathbf{A}\|_2^2 = \sigma_1^2$. We will sometimes use the notation $\|\mathbf{A}\|_\xi$ to indicate that an expression holds for both $\xi = 2$ or $\xi = \text{F}$. For any two matrices \mathbf{X} and \mathbf{Y} , $\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_\text{F} \leq \sqrt{\text{rank}(\mathbf{X})} \|\mathbf{X}\|_2$; $\|\mathbf{X}\mathbf{Y}\|_\text{F} \leq \|\mathbf{X}\|_\text{F} \|\mathbf{Y}\|_2$; $\|\mathbf{X}\mathbf{Y}\|_\text{F} \leq \|\mathbf{X}\|_2 \|\mathbf{Y}\|_\text{F}$. These are stronger variants of the standard submultiplicativity property $\|\mathbf{X}\mathbf{Y}\|_\xi \leq \|\mathbf{X}\|_\xi \|\mathbf{Y}\|_\xi$ and we will refer to them as spectral submultiplicativity. It follows that, if \mathbf{Q} is orthonormal, then $\|\mathbf{Q}\mathbf{X}\|_\xi \leq \|\mathbf{X}\|_\xi$ and $\|\mathbf{Y}\mathbf{Q}^\top\|_\xi \leq \|\mathbf{Y}\|_\xi$. Finally,

Lemma 15 (matrix-Pythagoras). *Let \mathbf{X} and \mathbf{Y} be two $n \times d$ matrices. If $\mathbf{X}\mathbf{Y}^\top = \mathbf{0}_{n \times n}$ or $\mathbf{X}^\top \mathbf{Y} = \mathbf{0}_{d \times d}$, then $\|\mathbf{X} + \mathbf{Y}\|_\xi^2 \leq \|\mathbf{X}\|_\xi^2 + \|\mathbf{Y}\|_\xi^2$.*

A.1 Sparsification Results

We now state two recent results on matrix sparsification ((Boutsidis, 2011, Lemmas 71 and 72, p. 132), (Boutsidis et al., 2011)) using our notation.

Lemma 16 (Spectral Sparsification). *Let $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ and $\Psi \in \mathbb{R}^{n \times \ell_2}$ with respective ranks $\rho_\mathbf{Y}$, and ρ_Ψ . Given $r > \rho_\mathbf{Y}$, there exists a deterministic algorithm that runs in time $T_{\text{SVD}}(\mathbf{Y}) + T_{\text{SVD}}(\Psi) + O(rn(\rho_\mathbf{Y}^2 + \rho_\Psi^2))$ and constructs sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$ satisfying:*

$$\text{rank}(\mathbf{D}\mathbf{S}\mathbf{Y}) = \text{rank}(\mathbf{Y}); \quad \|(\mathbf{D}\mathbf{S}\mathbf{Y})^+\|_2 < \frac{1}{1 - \sqrt{\rho_\mathbf{Y}/r}} \|\mathbf{Y}^+\|_2; \quad \|\mathbf{D}\mathbf{S}\Psi\|_2 < \left(1 + \sqrt{\frac{\rho_\Psi}{r}}\right) \|\Psi\|_2.$$

If $\Psi = \mathbf{I}_n$, the running time of the algorithm reduces to $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_\mathbf{Y}^2)$. We write $[\mathbf{D}, \mathbf{S}] = \text{MultipleSpectralSampling}(\mathbf{Y}, \Psi, r)$ to denote such a deterministic procedure.

Lemma 17 (Spectral-Frobenius Sparsification). *Let $\mathbf{Y} \in \mathbb{R}^{n \times \ell_1}$ and $\Psi \in \mathbb{R}^{n \times \ell_2}$ with respective ranks $\rho_\mathbf{Y}$, and ρ_Ψ . Given $r > \rho_\mathbf{Y}$, there exists a deterministic algorithm that runs in time $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_\mathbf{Y}^2 + \ell_2 n)$ and constructs sampling and rescaling matrices $\mathbf{S} \in \mathbb{R}^{r \times n}$, $\mathbf{D} \in \mathbb{R}^{r \times r}$ satisfying:*

$$\text{rank}(\mathbf{D}\mathbf{S}\mathbf{Y}) = \text{rank}(\mathbf{Y}); \quad \|(\mathbf{D}\mathbf{S}\mathbf{Y})^+\|_2 < \frac{1}{1 - \sqrt{\rho_\mathbf{Y}/r}} \|\mathbf{Y}^+\|_2; \quad \|\mathbf{D}\mathbf{S}\Psi\|_\text{F} \leq \|\Psi\|_\text{F}.$$

If $\Psi = \mathbf{I}_n$, the running time of the algorithm reduces to $T_{\text{SVD}}(\mathbf{Y}) + O(rn\rho_\mathbf{Y}^2)$. We write $[\mathbf{D}, \mathbf{S}] = \text{MultipleFrobeniusSampling}(\mathbf{Y}, \Psi, r)$ to denote such a deterministic procedure.

B Algorithms

Input: $A \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{b} \in \mathbb{R}^n$, and $r > k + 1$.

Output: sampling matrix \mathbf{S} and rescaling matrix \mathbf{D} .

- 1: Compute the SVD of $\mathbf{Y} = [\mathbf{A}, \mathbf{b}]$. Let $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{n \times \ell}$, $\Sigma \in \mathbb{R}^{\ell \times \ell}$ and $\mathbf{V} \in \mathbb{R}^{d \times \ell}$, with $\ell \leq k + 1$ (the rank of \mathbf{Y}).
- 2: **Return** $[\Omega, \mathbf{S}] = \text{SimpleSampling}(\mathbf{U}, r)$ (see Lemma 2)

Algorithm 1: Deterministic coreset construction for constrained linear regression.

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$.

Output: sampling matrix \mathbf{S} and rescaling matrix \mathbf{D} .

- 1: Compute the SVD of \mathbf{A} : $\mathbf{A} = \mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^T$, where $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$, $\Sigma_\mathbf{A} \in \mathbb{R}^{k \times k}$, and $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{d \times k}$; compute $\mathbf{E} = \mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^T\mathbf{B} - \mathbf{B}$.
- 2: **return** $[\mathbf{S}, \mathbf{D}] = \text{MultipleSpectralSampling}(\mathbf{U}_\mathbf{A}, \mathbf{E}, r)$ (see Lemma 16)

Algorithm 2: Deterministic coresets for multiple regression in spectral norm.

Input: $\mathbf{A} \in \mathbb{R}^{n \times d}$ of rank k , $\mathbf{B} \in \mathbb{R}^{n \times \omega}$, and $r > k$.

Output: sampling matrix \mathbf{S} and rescaling matrix \mathbf{D} .

- 1: Compute the SVD of \mathbf{A} : $\mathbf{A} = \mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^T$, where $\mathbf{U}_\mathbf{A} \in \mathbb{R}^{n \times k}$, $\Sigma_\mathbf{A} \in \mathbb{R}^{k \times k}$, and $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{d \times k}$; compute $\mathbf{E} = \mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^T\mathbf{B} - \mathbf{B}$.
- 2: **return** $[\mathbf{S}, \mathbf{D}] = \text{MultipleFrobeniusSampling}(\mathbf{U}_\mathbf{A}, \mathbf{E}, r)$ (see Lemma 17)

Algorithm 3: Deterministic coresets for multiple regression in Frobenius norm.

C Technical Proofs

Proof. (Theorem 4) We first construct \mathbf{D} and \mathbf{S} via Theorem 1 applied to \mathbf{A} and \mathbf{b}_{avg} . The running time is $O(n\omega)$ (the time needed to compute \mathbf{b}_{avg}) plus the running time of Theorem 1. The result is immediate from the following derivation:

$$\begin{aligned}
\|\mathbf{A}\tilde{\mathbf{X}}_{opt} - \mathbf{B}\|_{\mathbf{F}}^2 &\stackrel{(a)}{=} \omega\|\mathbf{A}\tilde{\mathbf{x}}_{opt} - \mathbf{b}_{avg}\|^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|^2 \\
&\stackrel{(b)}{\leq} \left(1 + O\left(\sqrt{k/r}\right)\right)^2 \omega\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}_{avg}\|^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|^2 \\
&\leq \left(1 + O\left(\sqrt{k/r}\right)\right)^2 \left(\omega\|\mathbf{A}\mathbf{x}_{opt} - \mathbf{b}_{avg}\|^2 + \sum_{i=1}^{\omega} \|\mathbf{b}_{avg} - \mathbf{B}^{(i)}\|^2 \right) \\
&\stackrel{(a)}{=} \left(1 + O\left(\sqrt{k/r}\right)\right)^2 \|\mathbf{A}\mathbf{X}_{opt} - \mathbf{B}\|_{\mathbf{F}}^2.
\end{aligned}$$

(a) follows by Lemma 3; (b) follows because $\tilde{\mathbf{x}}_{opt}$ is the output of a coresnet regression as in Theorem 1. Finally, $r > k + 1$ implies that $\left(1 + O\left(\sqrt{k/r}\right)\right)^2 = 1 + O\left(\sqrt{k/r}\right)$. \blacksquare

Proof. (Lemma 9) To simplify notation, let $\mathbf{W} = \mathbf{D}\mathbf{S}$. Using the SVD of \mathbf{A} , $\mathbf{A} = \mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^\mathbf{T}$, we get:

$$\|\mathbf{B} - \mathbf{A}\tilde{\mathbf{X}}_{opt}\|_{\xi}^2 = \|\mathbf{B} - \mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^\mathbf{T}(\mathbf{W}\mathbf{U}_\mathbf{A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^\mathbf{T})^+\mathbf{W}\mathbf{B}\|_{\xi}^2 = \|\mathbf{B} - \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\mathbf{B}\|_{\xi}^2,$$

where the last equality follows from properties of the pseudo-inverse and the fact that $\mathbf{W}\mathbf{U}_\mathbf{A}$ is a full-rank matrix. Using $\mathbf{B} = \left(\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\mathbf{T} + \mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\right)\mathbf{B}$, we get

$$\begin{aligned}
\|\mathbf{B} - \mathbf{A}\tilde{\mathbf{X}}_{opt}\|_{\xi}^2 &= \|\mathbf{B} - \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\left(\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\mathbf{T} + \mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\right)\mathbf{B}\|_{\xi}^2 \\
&= \|\mathbf{B} - \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\mathbf{U}_\mathbf{A}\mathbf{U}_\mathbf{A}^\mathbf{T}\mathbf{B} + \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\mathbf{B}\|_{\xi}^2 \\
&\stackrel{(a)}{=} \|\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\mathbf{B} + \mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\mathbf{B}\|_{\xi}^2 \\
&\stackrel{(b)}{=} \|\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\mathbf{B}\|_{\xi}^2 + \|\mathbf{U}_\mathbf{A}(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\mathbf{B}\|_{\xi}^2.
\end{aligned}$$

(a) follows from the assumption that the rank of $\mathbf{W}\mathbf{U}_\mathbf{A}$ is equal to k and thus $(\mathbf{W}\mathbf{U}_\mathbf{A})^+\mathbf{W}\mathbf{U}_\mathbf{A} = \mathbf{I}_k$ and (b) follows by matrix-Pythagoras (Lemma 15). To conclude, we use spectral submultiplicativity on the second term and the fact that $\mathbf{U}_\mathbf{A}^\perp(\mathbf{U}_\mathbf{A}^\perp)^\mathbf{T}\mathbf{B} = \mathbf{A}\mathbf{X}_{opt} - \mathbf{B}$. \blacksquare